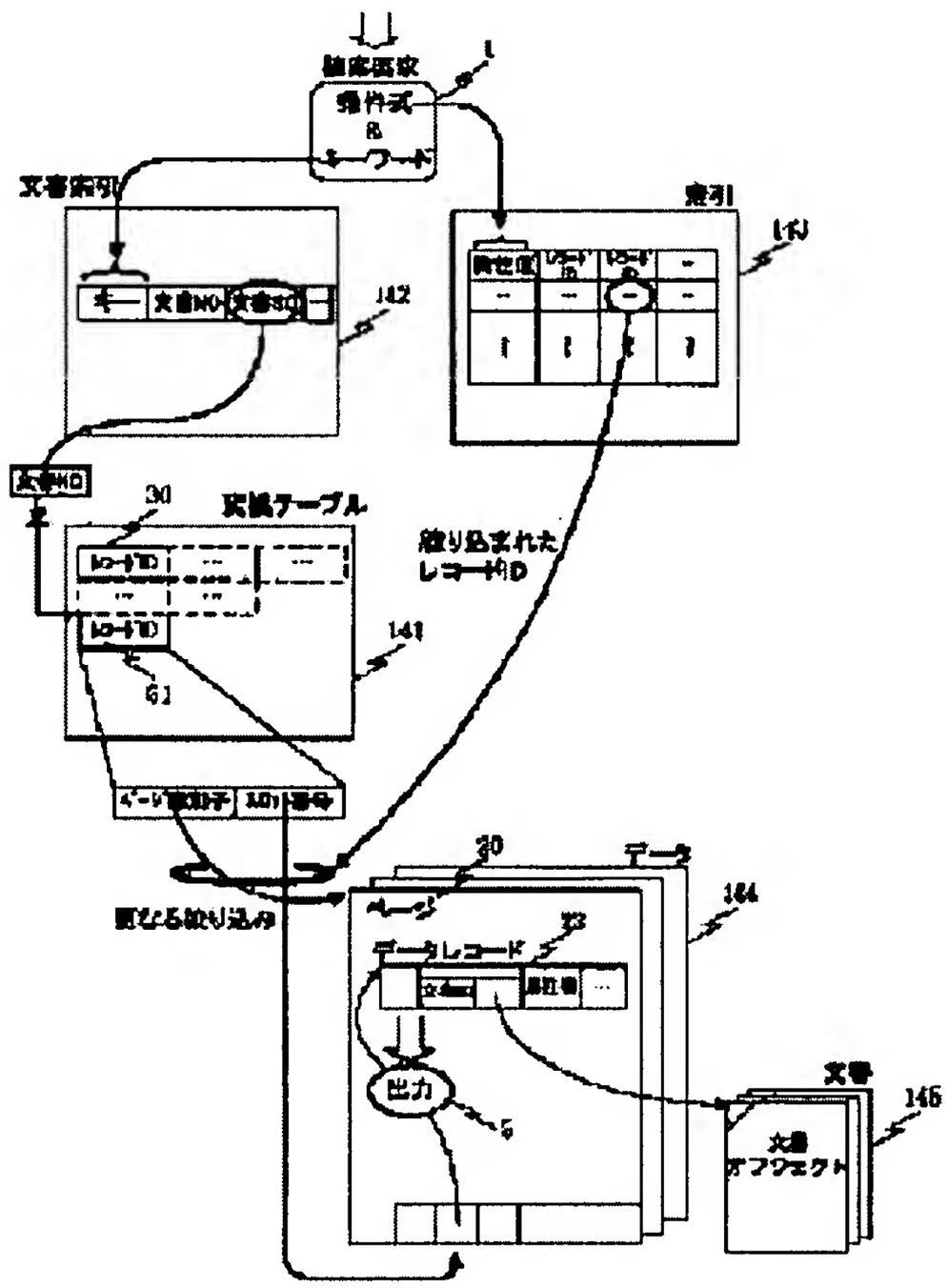
METHOD AND SYSTEM FOR MANAGING DATA BASE HAVING DOCUMENT RETRIEVAL FUNCTION

Patent number: JP9305622
Publication date: 1997-11-28
Inventor: HARA NORIHIRO; KAWAMURA NOBUO; KITAMURA KENICHI
Applicant: HITACHI LTD; HITACHI SOFTWARE ENG
Classification:
- international: ***G06F17/30;*** **G06F17/30;** (IPC1-7): G06F17/30
- european:
Application number: JP19960117311 19960513
Priority number(s): JP19960117311 19960513

Abstract of **JP9305622**

PROBLEM TO BE SOLVED: To reduce the capacity of document indexes and to improve the efficiency of retrieval corresponding to the request of inquiry with document retrieval conditions by uniquely recognizing stored documents and using a smaller sized 'document number' than a line identifier for document indexing. SOLUTION: While receiving a retrieval request 1 from the source of inquiry with a conditional expression and a keyword corresponding to the attribute value of data and referring to a document index 142 prepared corresponding to a document 145 based on the keyword, the document number of a document object containing the keyword is possessed. A record identifier 51 of entry of a conversion table 141 corresponding to the document No., is possessed. The document object containing the keyword of the retrieval request 1 is related through data 144 to the line of that conversion table 141. While using an index 143 prepared corresponding to the attribute value of data contained in the conditional expression of the retrieval request 1, the record ID of the line coincident with the conditional expression is possessed and the cluster of record ID is narrowed down while using the record ID provided from the index 143.



(19)日本国特許庁（ＪＰ）　(12)**公開特許公報**（Ａ）　(11)特許出願公開番号

特開平9－305622

(43)公開日　平成９年(1997)11月28日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ　17/30			Ｃ０６Ｆ　15/40	３７０Ａ	
			15/403	３５０Ａ	

審査請求　未請求　請求項の数４　ＯＬ　（全 15 頁）

(21)出願番号　特願平8－117311

(22)出願日　平成８年(1996) ５月13日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(71)出願人　000233055
日立ソフトウエアエンジニアリング株式会社
神奈川県横浜市中区尾上町６丁目81番地

(72)発明者　原　憲宏
神奈川県川崎市幸区鹿島田890番地の12
株式会社日立製作所情報・通信開発本部内

(74)代理人　弁理士　小川　勝男

最終頁に続く

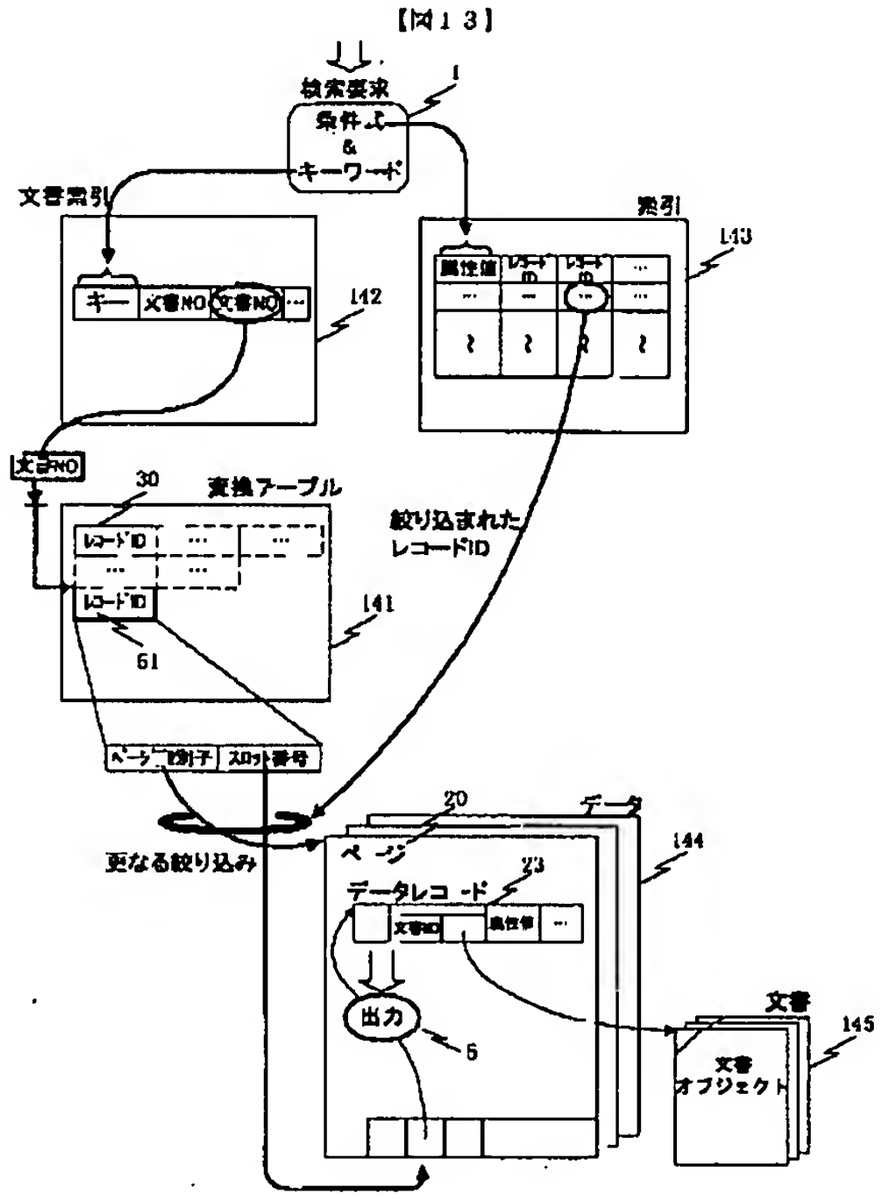
(54)【発明の名称】　文書検索機能を有するデータベース管理方法およびシステム

(57)【要約】

【課題】　テーブルおよびそのテーブルに関連付けられた文書を格納したデータベースに対して、効率の高い文書検索機能付きリレーショナルデータベース管理方法を提供すること、かつ文書検索の際に用いる文書索引の容量を削減することを課題とする。

【解決手段】　検索要求において文書検索条件に合致した文書145の文書番号集合を取得し、文書登録に際しては登録文書の文書番号を登録する文書索引142と、文書番号と対応するデータ144を一意に識別するデータ識別子とを関連付け、検索処理において文書索引142を参照することにより取得された文書番号集合を対応するデータ識別子集合に変換する変換テーブル141とからなる。

【図１３】



【特許請求の範囲】
【請求項１】計算機を用いて、文書オブジェクト、および前記文書オブジェクトと関連付けられたデータオブジェクトを管理するデータベース管理方法において、
ａ）入力手段から入力された検索要求に含まれるキーワードを含む少なくとも１つの文書の番号に対応する少なくとも１つの第１のレコード識別子を抽出し、
ｂ）前記検索要求に含まれる条件式を満たす属性値に対応する少なくとも１つのデータレコードの第２のレコード識別子を抽出し、
ｃ）前記第２のレコード識別子に合致する前記第１のレコード識別子を選択し、　ｄ）前記ステップｃ）で選択されたレコード識別子に対応するデータレコード、及び前記データレコードと関連付けられた文書オブジェクトを抽出することを特徴とする文書検索機能を有するデータベース管理方法。
【請求項２】前記ステップａ）は、
ａ１）前記キーワードによって文書索引を検索して少なくとも１つの文書番号を抽出し、
ａ２）前記抽出された文書番号に対応する格納位置に格納された前記第１のレコード識別子を変換テーブルから抽出することを特徴とする請求項１記載の文書検索機能を有するデータベース管理方法。
【請求項３】文書オブジェクト、および前記文書オブジェクトと関連付けられたデータオブジェクトを管理するデータベース管理システムは、
ａ）入力手段から入力された検索要求に含まれるキーワードを含む少なくとも１つの文書の番号に対応する少なくとも１つの第１のレコード識別子を抽出する手段、
ｂ）前記検索要求に含まれる条件式を満たす属性値に対応する少なくとも１つのデータレコードの第２のレコード識別子を抽出する手段、
ｃ）前記第２のレコード識別子に合致する前記第１のレコード識別子を選択する手段、及び
ｄ）前記選択されたレコード識別子に対応するデータレコード、及び前記データレコードと関連付けられた文書オブジェクトを抽出する手段を有することを特徴とする文書検索機能を有するデータベース管理システム。
【請求項４】文書オブジェクト、および前記文書オブジェクトと関連付けられたデータオブジェクトを管理するための、計算機で読み出し可能な記録媒体に格納されたデータベース管理方法のプログラムは、
ａ）入力手段から入力された検索要求に含まれるキーワードを含む少なくとも１つの文書の番号に対応する少なくとも１つの第１のレコード識別子を抽出し、
ｂ）前記検索要求に含まれる条件式を満たす属性値に対応する少なくとも１つのデータレコードの第２のレコード識別子を抽出し、
ｃ）前記第２のレコード識別子に合致する前記第１のレコード識別子を選択し、　ｄ）前記ステップｃ）で選択されたレコード識別子に対応するデータレコード、及び前記データレコードと関連付けられた文書オブジェクトを抽出することを特徴とする文書検索機能を有するデータベース管理方法プログラム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、文書検索機能を有するデータベース管理方法およびデータベース管理システムに関する。
【０００２】
【従来の技術】文書情報の蓄積、再利用の重要性に伴い、売り上げデータ等と同じように文書情報そのものをテーブル形式のデータベースとして格納管理し、任意の文字列を入力して種々の条件で文書検索を行うデータベース管理システムが望まれている。このため、データベース内の論理的構造がテーブルの、行（ロー）、列（カラム）から構成されるリレーショナルデータベースにおいて、列に含まれるデータ型として文書格納のための文書型を対応させ、その文書型列に対する検索処理という形態で上記要求に応えるシステムが提供されている。それらデータベース管理システムでは、行の格納形態であるレコード内のカラム対応部分に文書格納領域へのポインタを格納することによって、あたかも文書実体が行内に存在しているかのように見せている。
【０００３】一方、一般の文書検索システムでは、大量の文書を高速に検索するために、前処理で索引を作成する。文字列を入力して検索条件を満たす文書を確定するために、文書本体をすべてアクセスするのでは検索効率が悪いからである。文書に含まれる文字列を文書から切り出して、その文字列をキーとして索引を構成し、検索時にその索引を用いることにより、文書本体へのアクセス無しに効率的に検索条件を満たす文書を確定する。その場合、文書からの文字列の切り出し方、すなわちキーの構成方法が索引の容量に大きく影響する。切り出し文字列として一文字を用いると、索引キーから文書へのポインタ付けが漏れなく行われ、検索条件判定の際に文書本体をアクセスする必要がなくなる。しかし、索引容量が莫大になり、結局検索効率の向上は十分には得られない。また、切り出し文字列長を増やしていくと、文書へのポインタ付けが荒くなり、それを補うため文書本体をアクセスしなければならなくなってしまう。検索効率を向上させるためには、文書へのポインタ付けの漏れがないように索引を作成し、かつその容量を減らすことが重要となる。
【０００４】通常のリレーショナルデータベース管理システムでは、各行の列値をキーとする索引(インデクス)を用いて検索効率の向上を図っている。索引の構造としてＢ木構造を持つＢ木インデクスがよく用いられる。インデクスキーから行（ロー）への関連付けは、行（ロー）を一意に識別できる「行（ロー）識別子」によって

行われる。すなわち、インデクスはキーと行（ロー）識別子から成るインデクスエントリによって構成される。この「行（ロー）識別子」は、インデクスエントリの構成要素であるとともに検索結果集合に対する集合演算処理のような、リレーショナルデータベースに対するデータ処理の対象データである行の識別子として用いられる。「行（ロー）識別子」を用いて行本体への高速アクセスが可能なように、「行（ロー）識別子」は、データベースを格納するファイルのアクセス単位であるページのページ番号とページ内の行格納位置情報で構成されることが多い。

【０００５】

【発明が解決しようとする課題】従来、文書検索機能付きリレーショナルデータベース管理システムでは、文書索引の構成要素である文書へのポインタとして、Ｂ木インデクスと同様に、関連文書が属する行の「行（ロー）識別子」を用いていた。そのため、文書索引から検索条件を満たす「行（ロー）識別子」集合を取得し、これと他カラムに対する検索条件からの結果集合との集合演算等が可能となる。しかし、その場合、文書管理システムの文書索引内で管理される文書へのポインタ、または文書識別手段に比べ、一般にファイル全体の中の格納位置を示す「行（ロー）識別子」のサイズの方が大きいため、文書索引の容量が莫大になり、結果的に検索効率の低下を招いてしまうという問題があった。

【０００６】本発明は、これらの問題を解決するため、格納文書を一意に認識でき、行識別子よりサイズの小さい「文書番号」を文書索引に用いることにより、文書索引の容量を削減すると共に文書検索条件を伴う問合せ要求に対する検索効率の高い文書検索機能付きリレーショナルデータベース管理方法を提供することを目的としている。

【０００７】

【課題を解決するための手段】上記目的を達成するために、本発明におけるデータベース管理システムは、以下の構成を有する。

【０００８】データベースの文書に文書オブジェクトを登録する際に、当該文書オブジェクトを一意に識別して、文書索引に用いられる文書番号の割当てを行う手段を有する文書番号管理部と、検索要求の際に、文書に対応して作成された文書索引に基づいて、文書検索条件に合致した文書オブジェクトの文書番号を集合の形で取得する手段と、文書登録の際に書索引を管理する手段とを有する文書索引管理部と、文書番号管理部によって格納文書中の各文書オブジェクトに割り当てられた文書番号と、文書オブジェクトに関連付けられているデータ中の各データオブジェクトを一意に識別するためのデータ識別子とを関連付ける変換テーブルを設け、検索操作の際に文書索引管理部により取得された文書番号集合を関連付けられているデータ識別子に変換する手段とを有する変換テーブル管理部と、を備える。

【０００９】

【発明の実施の形態】以下、図を用いて本発明の実施の一形態を詳細に説明する。本発明は、図１４に示す計算機システムで実施される。図4の計算機システムは、中央処理装置（ＣＰＵ）100、入出力端末（ＶＤＴ）200、及びディスク装置300からなり、ディスク装置には、後述するデータベース（ＤＢ）4、及び本発明による処理手順を実行するプログラム500が格納されている。プログラム500は、ＣＰＵの主記憶に読み込まれて実行される。

【００１０】まず、図１３の本発明の概念図を説明する。本発明のデータベース管理システムでは、データの属性値に対する条件式およびキーワードを伴う問合せ元からの検索要求１を受け付けた際、キーワードを元に文書145に対応して作成された文書索引142を参照し、そのキーワードを含む文書オブジェクトの文書番号(文書NO.)(群)を取得する。そして、文書NO.に対応する変換テーブル141のエントリ(文書NO.から算出される格納位置)内に記憶してあるレコード識別子51(レコードID)(群)を取得する。レコード識別子51は、データ144におけるデータレコード23の格納位置を示す情報であり、データ内のページの識別子とページ内の格納位置を格納したスロットの番号とで構成される。取得したレコード識別子51を持つ変換テーブル141の行には、検索要求1のキーワードを含む文書オブジェクトがデータ144を介して関連付けられている。

【００１１】また、検索要求１の条件式に含まれるデータの属性値に対応して作成された索引143を用いて、条件式に合致した行のレコードID(群)を取得する。ここで、変換テーブル141を参照して得られたレコードIDの集合を、先の索引143から得られたレコードIDを用いて絞り込む。

【００１２】ここで、絞り込まれた結果に含まれるレコードIDからページ20内に格納されているデータレコード23(テーブルの行の格納形態)をアクセスし、検索結果として文書オブジェクトへのポインタなどを出力する(5)。

【００１３】次に図１には、本発明のデータベース管理システムの構成図が示してある。図１に示すように、本発明のデータベース管理システムは、問合せ元からの問合せ要求1を受付けて解析し、問い合わせ要求に応じてデータベース４の検索処理および更新処理を行うデータベース処理部3から構成される。問合せ要求・結果処理部2は、利用者からの問合せ要求1を受付けて解析し(121)、問合せ要求に対応したデータ処理の実行をデータベース処理部3に要求し(122)、データベース処理部3から問い合わせ結果を処理して(123)、問合せ元に問合せ結果5を出力する。

【００１４】データベース処理部３は、問合せ要求・結

果処理部２からの要求に応じて、データベース４を検索あるいは更新し、その結果を問合せ要求・結果処理部２に返す。データベース４の検索あるいは更新処理を担当するのが、文書索引管理部131、索引管理部133、データ管理部134、変換テーブル管理部132、そして文書番号管理部135である。

【００１５】ここで、図１における矢印は、検索要求の際の処理の主な流れを示す。文書索引管理部131および索引管理部133を用いることにより効率よく検索結果を絞り込む。要求によっては、データ管理部134が検索結果として絞り込まれたデータを参照する。文書番号管理部は、文書登録の際に文書番号の割り当てを行う。

【００１６】本発明のデータベース管理システムで管理されるデータベース4は、データベース操作対象であるデータオブジェクトの集まりから成るデータ144、データ144に対応して作成された索引143、データ144のデータオブジェクトそれぞれに関連付けられた文書オブジェクトの集まりから成る文書145、文書145に対応して作成された文書索引142、そして上記文書145の文書オブジェクトとデータ144のデータオブジェクトとを論理的に結び付けるための変換テーブル141からなる。

【００１７】データ144のデータオブジェクトの一例として、リレーショナル管理システムにおけるデータモデルであるテーブルの構成要素である行が挙げられる。データ144のデータオブジェクト、すなわち本形態における行は、データベース４へのアクセス単位であるページ中に、データレコードという形態で格納される。そのデータレコードに対して、問合せ要求・結果処理部２の問合せ処理実行制御122の指示によって、格納/読み出し等を担当するのがデータベース処理部３のデータ管理部134である。検索高速化のためにしばしばデータに対して索引143を作成し、検索時参照する。その索引143の参照および更新処理を行うのが、索引管理部133である。

【００１８】本発明のデータベース管理システムでは、文書をテーブルの列の属性値として提供するに当たり、データ144と文書145とをそれぞれ別領域に格納して互いに関連付ける。その関連付けの手段が変換テーブル141である。また、文書高速検索手段として文書索引142を有し、文書索引の維持管理をデータベース処理部３内の文書索引管理部131が担当する。文書索引管理部131は、文書検索要求に際し、文書索引142を参照することにより文書検索条件に合致した文書に関する情報を取得する。文書145内の文書オブジェクトは、データベース格納時に文書番号管理部135によって割り当てられた文書オブジェクトを一意に認識するための文書番号によって識別される。文書索引管理部131は、文書オブジェクトに関する情報として文書番号を取得する。

【００１９】テーブルの行の識別手段として、ある条件等に合致した行集合に対して集合演算を施したり、特定行をアクセスしたりするために、行に対応する格納データレコードのレコード識別子（データ識別子）を用いる。データオブジェクトと文書オブジェクトの関連付けは、上記文書番号とレコード識別子を用いて行う。

【００２０】図２は、データ144の各ページ内におけるデータレコードの格納構造の一形態を示す図である。１つのページ20内には、複数のデータレコード23が格納可能であり、データレコード23のページ内の格納位置は、スロット21により指示される。スロット21の領域には指示するデータレコード23が格納されているページ20の先頭からの格納位置が記憶される。ページ制御情報22は、スロットの割当て状況などのスロット管理およびページ内領域管理を行うためのものである。データレコード23は、文書を属性値として持つ列(カラム)に対応する文書フィールド24を含む。文書フィールド24は、文書を一意に認識するための文書番号25および、文書本体をアクセスするためのポインタ（文書格納位置情報）26から成る。文書番号25は、データレコード23と文書145中の各文書オブジェクトとを論理的に結び付けるために用いられる。その文書番号25に対し、文書オブジェクトへのポインタ26は両者(データレコードと文書オブジェクト)を物理的に結び付けるために用いられる。

【００２１】図３は変換テーブル141に格納されている各レコード識別子の構成の一形態を示す図である。レコード識別子51は、データレコード(図２の23)が格納されるページ(図２の20)を一意に識別するためのページ識別子31と、ページ内のデータレコード格納位置を特定するためのスロット(図２の21)を示すスロット番号32から成る。スロット番号32は、ページ格納構造においてページ制御情報(図２の22)側から順次番号付けされる。図３では、「ページ識別子＋スロット番号」という構造を採っているが、「スロット番号＋ページ識別子」でもなんら問題はない。レコード識別子51を用いてデータレコードをアクセスする。データレコードへのアクセスは、このレコード識別子のページ識別子51を用いて格納ページをアクセスし、スロット番号32に対応するスロットに記録されているデータレコード格納位置を取得することによって高速に行われる。

【００２２】図４は、文書索引および索引の具体例を示す図である。図４のa)は、図１のデータベース4内の文書索引142(図１３の概念図にも記述)の詳細構成例である。また図４のb)は、図１のデータベース4内の索引143(図１３の概念図にも記述)の詳細構成例である。

【００２３】文書索引142の中には各インデクスキーワードに対応した複数の索引41が含まれる。ここで、先頭の“本”はインデクスキーワードであり、それに続く文書番号11、12、…、1nは、インデクスキーワード“本”を含む文書オブジェクトの文書番号である。同様に、“発”および“明”について図示のように登録されている。この構造によりどんなキーワードが検索列としてやってきても文書オブジェクト本体をアクセスすることな

しに検索条件に合致した文書番号を取得できる。
【００２４】索引143の中には、属性値とその属性値を持つ列(カラム)の行を示すレコードID(レコード識別子)(群)から成る索引エントリ42が記録されている。属性値を指定すると、容易にその属性値を持つレコードIDを取得することができる。ここでは、索引エントリはテーブル構造をとっているが、B木構造やハッシングを用いた構造でもよい。
【００２５】図５は、変換テーブルの一例を示す図である。これは、図１のデータベース4内の変換テーブル141(図１３の概念図にも記述)の詳細例である。本変換テーブルは、上記文書索引によって取得した検索条件に合致する文書番号を、リレーショナルデータベース管理システムが種々の演算において採用するレコード識別子(図３の30)に変換するためのものである。変換テーブル141は、変換テーブルエントリ51により構成される。本形態において変換テーブルエントリ51はレコード識別子(図３の30)から構成される。
【００２６】そして、テーブル141は、複数の変換テーブルエントリ51の格納位置を文書番号から計算により容易に特定できるような構造になっている。さらに具体的に述べると、文書番号を１から順に格納領域をインクリメンタルに割り当てることにより、そのシリアルな文書番号に対応するレコード識別子30が変換テーブルの対応するエントリにマッピングされるようにする。その結果、文書番号より変換テーブルのエントリをアクセスし、エントリに記録してある対応レコード識別子を取得できる。
【００２７】各変換テーブルエントリの構成要素がページ識別子とスロット番号からなるレコード識別子のみであり、文書番号やエントリ自身の情報などを必要としないことから、変換テーブルの容量を必要最小限に抑えることができる。さらに、文書索引内にレコード識別子を持つ場合、文書索引内には同一レコード識別子が大量に存在しそれがアクセス効率の低下を招く要因になることから、変換テーブルを参照し文書番号からレコード識別子に最終的に変換する方が効率よくアクセスすることができる。さらなる変換テーブルアクセス効率向上のため、変換テーブルはメモリに常駐させる方が望ましい。
【００２８】次に図２から図５で説明した一構成形態のもとで、図６および図７を用いてデータベースの検索処理について詳細に説明する。
【００２９】図６は、検索要求が問合せ元から入力された際の、データベース処理の詳細を示すフローチャートであり、図１における問合せ実行制御122以降の処理について示している。まず、ステップ601において、要求検索操作は文書索引を使用する検索であるかどうかを判定する(図１の122)。文書索引の使用不使用の指定は、図１の問合せ要求受付け・解析121において問合せ要求に含まれる検索条件により決定される。(ここで図１でのデータベース処理部3に制御が渡る。)ステップ601において文書索引使用指定の場合、ステップ602以降に進み文書検索条件による検索実行を行う。文書索引の使用の指定がない場合、ステップ609に進み、索引による検索を行うかどうかの判断を行う。
【００３０】ステップ602に進んだ場合、図１の文書索引管理部131が以下の処理を行う。文書索引をアクセスし(ステップ602)、文書検索条件を満たす文書番号集合を取得する(ステップ603)。次に、取得した文書番号の集合を対応するレコード識別子の集合に変換するために、文書番号一つ一つを評価する。すなわち、ステップ604において、文書番号集合に要素すなわち文書番号が存在するかを判定する。文書番号が存在しない場合、文書番号集合はすべてレコード識別子集合に変換完了したとみなし、ステップ609に進む。
【００３１】集合要素である文書番号が存在する場合、ステップ605以降により一文書番号の変換処理を行う。変換処理は、図１の変換テーブル管理を通して行う。ステップ605において文書番号集合から一文書番号を取り出す。そして、文書番号から変換テーブルの対応するエントリの格納位置を算出し、変換テーブルの対応するエントリをアクセスする(ステップ606)。そして、変換テーブルのエントリからレコード識別子を取得する(ステップ607)。取得したレコード識別子を変換結果としてレコード識別子集合１に追加する(ステップ608)。そして、ステップ604に戻り、残りの文書番号の変換を続行する。
【００３２】文書番号集合の変換処理がすべて終了後、または文書索引による検索処理を行わなかった場合、ステップ609において、索引を使った検索を行うかどうかを判定する。索引を使った検索を行う場合、図１の索引管理部133に処理制御が渡り以下の処理を行う。ステップ610に進み索引をアクセスし、検索条件を満たすレコード識別子集合２を取得する(ステップ611)。ステップ609において、索引を使用しないと判断した場合、ステップ612に進む。ステップ612において、検索条件の組合せによる集合演算を行う。具体的には、文書に対する検索条件と索引を使うような検索条件のAND条件で問合せ要求がなされている場合は、レコード識別子集合１とレコード識別子集合２の積集合を結果レコード識別子集合とする。
【００３３】また、OR条件の問合せ集合の場合には、レコード識別子集合１とレコード識別子集合２の和集合を結果レコード識別子集合とする。どちらかの条件のみの場合はレコード識別子集合の集合演算は行わず、そのまま結果レコード識別子集合とする。その後、ステップ613において、結果レコード識別子集合を用いて要求に応じレコードをアクセスし(ステップ613)、結果として問合せ元に返す(ステップ614)。
【００３４】図７は、本発明の検索動作説明図を示して

いる。これは、図６のフローチャートに従って説明した検索時の具体例である。データベース4には、データ144および文書145として、図７に示す「著者」列および「文書」列（文書型）を持つテーブルが格納管理されている。問合せ元が問合せ要求1として、「著者＝HARA」かつ「"データベース"を含む」行の検索を要求する。処理122において上記問合せ要求を受付けて解析し、アクセス手段の決定を行う。本具体例では、「著者」列に索引が定義され、文書索引が用意されているので、索引および文書索引を用いて検索処理を行うことを決定する。

【００３５】そして、処理122において決定されたアクセス手段に従って検索処理を以下のように制御する。まず、文書索引管理部131において文書索引142をアクセスし、検索条件合致文書番号(文書番号１、文書番号２)を取得する(図６のステップ603に相当)。そして、変換テーブル管理部132において、変換テーブル141を参照し、先に取得した文書番号集合をレコード識別子集合（レコード識別子n、レコード識別子m）に変換する(図６のステップ605からステップ608に相当)。次に、検索条件「著者＝HARA」より、索引管理部133において索引143をアクセスし、検索条件合致レコード識別子（レコード識別子m、レコード識別子k）を取得する。検索結果処理部123において、上記結果レコード識別子集合をマージし(本実施例の場合、積集合を求める)(図６のステップ612に相当)、最終結果レコード識別子mを取得し、検索結果709として問合せ元に返す(図６のステップ614に相当)。

【００３６】以上によって、文書が格納された列「文書」に対する文書検索条件を含む検索操作を、文書番号からレコード識別子への容易な変換を用いて、他の列に作成されている索引を利用するのと同じ要領で実行することができた。また、索引と併用することで、結果集合の縛り込みが効率的に行えた。ここでは、「文書」列以外の列に作成されている一索引の利用例を示したが、検索条件によっては複数索引を用いてもよい。また、データベースへのI/O数等を加味して最適化を図り、適切な索引を組合せて使用するようにしてもよい。

【００３７】次に、図８および図９を用いてデータベースへの登録操作について詳細に説明する。図９は、本発明の登録操作フローチャートであり、図６同様に図１における問合せ実行制御122以降の処理について示している。

【００３８】新規データおよび新規文書の登録の際に、まずデータベース処理部３では、ステップ801にて新規文書番号の割り付けを行う。これは、図１の文書番号管理部135が行う。文書番号の管理方法の一形態として、文書番号を「採番カウンタ」で管理し、新規文書の登録において「採番カウンタ」に＋１した値を文書番号として割り当てる。その際「採番カウンタ」の値は＋１する。ここで、文書番号(採番カウンタ)を実現するために必要なビット数（サイズ）は、レコード識別子を構成するページ識別子およびスロット番号を実現するためのビット数（サイズ）よりも小さい。これは、レコード識別子の割当てがまばらになるのに対して、文書番号は常に順番に割り当てられることからも分かる。

【００３９】次に、データベース4に新規文書オブジェクトを格納する(ステップ802)。先程の新規文書番号および新規文書格納位置を用いて、新規データレコードを作成する(ステップ803)。新規データレコードの作成に当たり、図２を用いて説明したデータレコード23の文書フィールド24の文書番号25および文書オブジェクトへのポインタ26（文書オブジェクト格納位置）を設定する。そして、新規データレコードを格納するためのページを決定し(ステップ804)、格納ページ内のページ制御情報から新規データレコードのためのスロットを割り当ててもらい(ステップ805)、新規データレコードをページ内に格納する(ステップ806)。格納ページおよびスロット番号決定時、レコード識別子が確定する。

【００４０】データレコード格納後、文書列以外の列に索引が存在するかを判定し(ステップ807)、存在する場合その索引のメンテナンスをレコード識別子を用いて行う(ステップ808)。さらに、ステップ809において、ステップ801で割当てた文書番号を用いて文書索引のメンテナンスを行う。文書番号から変換テーブルエントリの位置を算出し(ステップ810)、変換テーブルエントリにステップ805までに確定したレコード識別子を設定する(ステップ811)。

【００４１】ここでは、索引のメンテナンス処理の後に、文書索引のメンテナンスを行っているが、文書番号の割当ておよびレコード識別子の確定が完了していさえすれば、索引および文書索引のメンテナンスの順序に制約はない。もちろん両メンテナンス処理は処理高速化のため並列に実行することが望ましい。また、変換テーブルエントリのメンテナンスも文書番号および対応レコード識別子が確定した段階で行って構わない。

【００４２】図９は、本発明の登録操作説明図を示している。これは、図８のフローチャートに従って説明した登録時の具体例である。データベース４には、図７と同様に「著者」列および「文書」列を持つテーブルがデータ144および文書145として格納管理されている。問合せ元が問合せ要求1として、「著者＝NISHI」であり文書オブジェクト「…。インターネットは、…。」を伴う新規データの登録を要求する。処理121において上記問合せ要求を受付けて解析する。そして、処理122において登録処理を以下のように制御する。

【００４３】まず、データ管理部134において文書格納を行うが、それに先立ち文書番号管理部135において文書番号の割り付けを行い、「文書番号４」を取得する(図８のステップ801に相当)。次に新規文書オブジェクトをデータベース４に格納し(図８のステップ802に相当)、そのポインタと「文書番号４」を用いてデータレ

コードの格納を行う(図８のステップ806に相当)。その際、「レコード識別子p」が確定される。「著者」列に索引が作成されていることから、インデクスキー「NISHI」および「レコード識別子p」を用いて索引管理部133において索引143のメンテナンス処理を行う(図８のステップ808に相当)。それとともに、「文書番号４」を用いて文書索引131において文書索引142のメンテナンス処理を行う(図8のステップ809に相当)。またそれとともに、変換テーブル管理部132において、「文書番号４」から変換テーブルエントリ位置を算出し、エントリに「レコード識別子p」を設定することにより、新エントリ設定を完了する(図８のステップ811に相当)。

【００４４】以上によって、「著者＝NISHI」を含むデータレコードと新規文書オブジェクト「…。インターネットは、…。」とを関連付けて登録することができる。

【００４５】次に、図１０および図１１を用いてデータベースからの削除操作について詳細に説明する。図１０は、本発明の削除操作フローチャートであり、図６と同様に図１における問合せ実行制御122以降の処理について示している。データおよびそれに関連する文書の削除の際に、まず、データベース処理部では、ステップ1001にてレコード識別子を用いて削除対象となっているデータレコードの削除を行う。その際、関連する文書オブジェクトの文書番号および文書格納ポインタを記憶しておく。

【００４６】次に、ステップ1002にて先に記憶しておいた文書格納ポインタを用いて文書オブジェクトの削除を行う。そして、ステップ1003において削除データレコードに関連する索引のメンテナンスを行う。すなわち、ステップ1003で索引が作成されているかどうかを判定する。索引がある場合、ステップ1004において列の値および削除対象レコード識別子を用いて索引メンテナンス(索引エントリの削除)を行い、ステップ1005に進む。索引が存在しない場合、そのままステップ1005に進む。次に、削除データレコードから記憶しておいた文書番号を用いて文書索引のメンテナンスを行う(ステップ1005)。さらに文書番号から変換テーブル位置を算出し(ステップ1006)、変換テーブルエントリ内のレコード識別子を初期化することにより、対応変換テーブルエントリを無効化する(ステップ1007)。

【００４７】図１１は、本発明の削除操作説明図を示している。これは、図１０のフローチャートに従って説明した削除時の具体例である。データベース４には、図７および図９と同様に「著者」列および「文書」列を持つテーブルがデータ144および文書145として格納管理されている。問合せ元が問合せ要求1として、「著者＝NISHI」であるデータ(行)の削除を要求する。処理121において上記問合せ要求を受付けて解析する。そして、処理122において削除処理を以下のように制御する。

【００４８】まず、データ管理部134において削除対象レコードの「レコード識別子p」を確定し、データレコードの削除を行う(図１０のステップ1001に相当)。そして、データレコードに格納されていた関連文書オブジェクトへのポインタを用い、対応する文書オブジェクトの削除を行う(図１０のステップ1002に相当)。図１１に示すように、データ144および文書145の削除対象レコードおよび文書オブジェクトを点線で示してある。

【００４９】削除データレコードの「レコード識別子p」、索引がはられている列の値、および削除データレコードに格納されていた関連文書オブジェクトの「文書番号４」を用いて、索引管理部133にて索引メンテナンス処理を、文書索引管理部131において文書索引メンテナンス処理を、さらに変換テーブル管理部132において対応エントリの初期化をそれぞれ行う(図１０のステップ1001、ステップ1004、ステップ1005にそれぞれ相当)。対応エントリの初期化において、「文書番号4」から変換テーブルの位置を算出しエントリ内の「レコード識別子p」を初期化する。

【００５０】以上によって、削除要求処理を完了する。図１１の流れからも分かるように、データレコード削除の後、関連する文書番号および削除文書格納ポインタ、索引キーが確定されているので、それらの処理を順次実行する理由はない。これは、並列処理による高速化を意味する。

【００５１】上述の実施形態では、文書番号割当て処理を文書番号管理部135が行っているが、データベースへの文書格納を行うためのデータ管理部134が行っても構わない。また、変換テーブルの実施形態に関しても、上述のようなエントリの配列構造ではなく、Ｂ木構造であってももちろん構わない。

【００５２】また、図１２に文書索引のさらなる一実施形態を示す。図１２は、ビットマップ・インデクスを用いた文書索引の例を示す図である。これは、図４における文書索引例とは別の実施形態である。図４の例では文書索引が文字列をキーとする文書番号リストより構成されていたのに対し、図１２の例では、文字列に対してビットマップ・インデクスを作成する。格納文書オブジェクトそれぞれに１ビットを割り当てる。また、各ビットは文書オブジェクト中の文字列に対応する。例えば、文字列が“本”のビットマップ・インデクスにおいて、１番目とｎ番目のビットが１であれば、１番目とｎ番目の文書オブジェクトに“本”が含まれていることを意味する(1200)。

【００５３】ビットマップ・インデクスを用いた場合、ビットマップの中の位置によって文書を識別する。最初のビットが文書番号１を指し、ｎ番目のビットが文書番号ｎを指す。本実施形態において、文書番号は１からシリアルに割当てられるため、ビットマップ・インデクスによる実現は容易であり、アクセスも効率的である。また、ビットマップ・インデクスを用いた場合、検索結果

は一時的にビットマップの形態をとることになる。これは、テーブルに複数の文書列が定義され、その列おのおのに検索条件が指定された問合せ要求が入力された場合、複数の文書索引を用いた検索結果のAND/OR演算は、レコード識別子に一旦変換しなくとも、ビットマップ同士のビット演算で効率良く実現できることを意味する。また、以上のことは、文書索引を用いた文書検索に縛られることなく、ビットマップ・インデクスを用いた列における高速検索への拡張も意味する。

【００５４】すなわち、文書属性を持たない列に対する索引をビットマップ・インデクスで構築した場合でも、本発明における変換テーブルを用いることにより、レコード識別子（行識別子）とビットマップ・インデクス内のビット位置との対応が容易にでき、かつ索引中にレコード識別子を持たないことから索引の容量も小さく抑えることができる。

【００５５】文書は多数の文字列を含むため、文書索引内には、同一文書番号が多数存在する。変換テーブルを導入することにより、文書番号として最小サイズのものを採用することができ、レコード識別子を用いて文書索引を構成するよりも、文書索引の容量を削減することができた。また、変換テーブルは、文書番号から対応レコード識別子を取得する目的のみに用い、種々の問合せよりに関してレコード識別子から文書番号への逆変換はフローチャートからも分かるように不要である。しかも、文書番号からレコード識別子への変換は容易に実行することができることから、文書を伴うデータの検索、登録、又は削除操作が効率良く行うことができる。

【００５６】本発明における一形態として、データオブジェクトと文書オブジェクトが１対１に対応する形態を説明したが、別形態として、データオブジェクトと文書オブジェクトとの関連は、多対１、１対多、多対多でも構わない。また、データオブジェクト新規格納時において、関連文書オブジェクトが決定あるいは格納されていなくともよい。その場合、データオブジェクトの文書フィールドに、文書オブジェクト未決定情報（具体的にはnull値）を記録しておけばよい。これらのことは、データオブジェクトと文書オブジェクトとが、ユティリティにおけるデーター括登録などにおいて独立した運用が可能であることを示している。文書オブジェクトの登録には一般にデータオブジェクトの格納に比べ多数のI/Oを伴うため、データオブジェクトのみを先に一括登録し、文書オブジェクトの登録は後回しにしておく、等という運用も容易である。

【００５７】

【発明の効果】以上説明したように、本発明によれば、格納文書の識別手段としてレコード識別子よりサイズの小さい「文書番号」を採用し、文書索引内において文書番号を用いて索引文字列との関連を管理し、検索時にその「文書番号」と関連する行の「レコード識別子」とを変換テーブルを用いて容易に「文書番号」から「レコード識別子」に変換し、レコード識別子に条件式を適用することにより、文書検索条件を伴う問合せ要求に対し効率よく検索することができる。さらに、索引を併用することにより、レコード識別子の絞り込みを効率的に行うことができる。

【図面の簡単な説明】

【図１】本発明の原理構成図である。

【図２】データレコードの格納構造の一形態を示す図である。

【図３】レコード識別子の構成の一形態を示す図である。

【図４】文書索引の一形態を示す図である。

【図５】本発明の文書番号レコード識別子変換テーブル構造の一例を示す図である。

【図６】発明の検索操作フローチャートである。

【図７】本発明の検索操作説明図である。

【図８】本発明の登録操作フローチャートである。

【図９】本発明の登録操作説明図である。

【図１０】本発明の削除操作フローチャートである。

【図１１】本発明の削除操作説明図である。

【図１２】ビットマップ・インデクスを用いた文書索引の例である。

【図１３】本発明の概念図である。

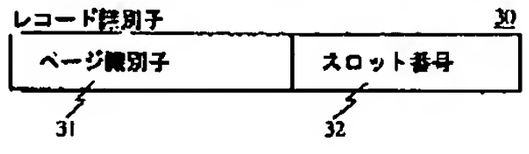
【図１４】本発明を実施する計算機システムの構成図である。

【符号の説明】

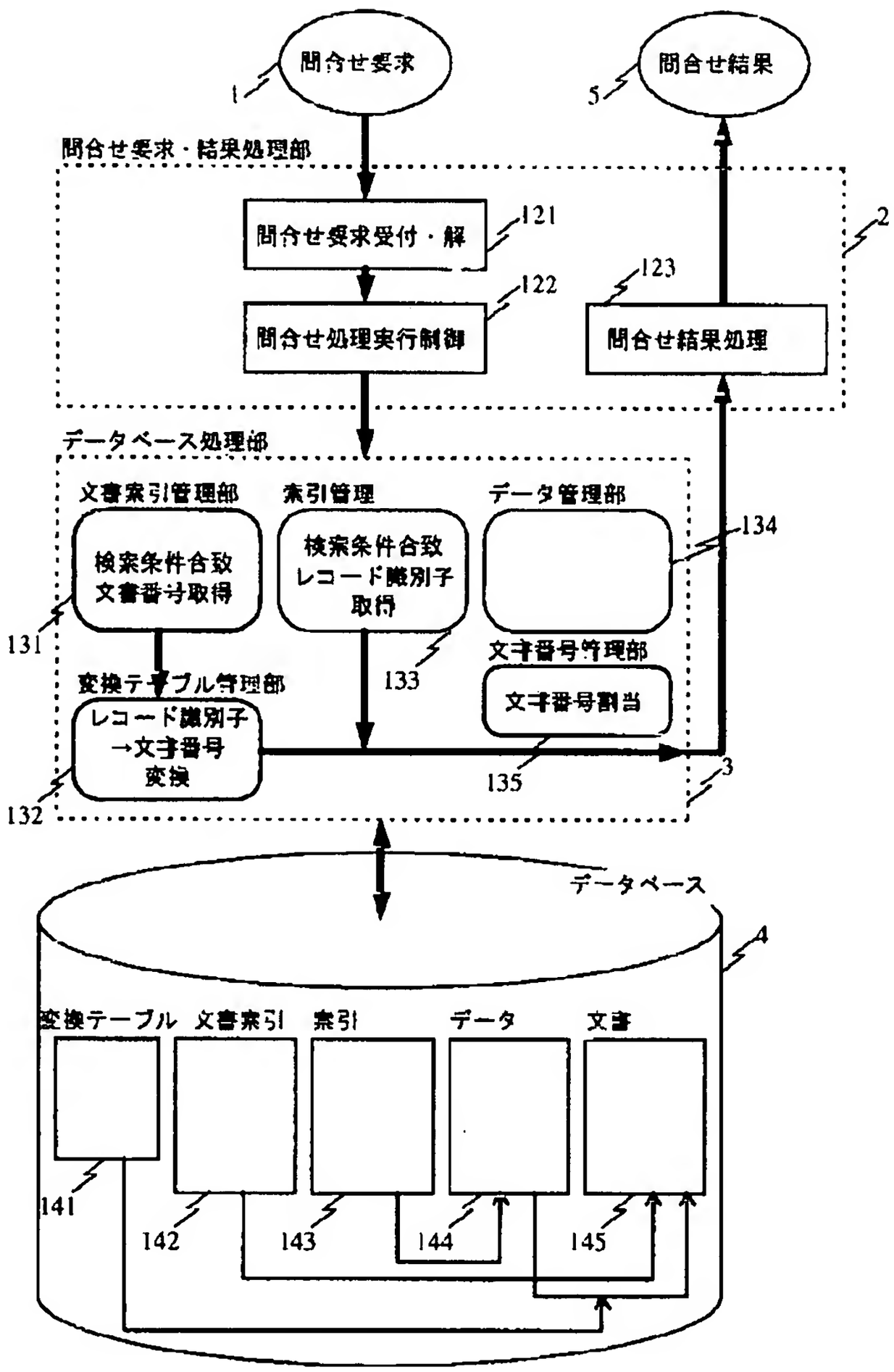
１：問合せ要求、１３１：文書索引管理部、１３２：変換テーブル管理部、１３３：索引管理部、１３４：データ管理部、１３５：文書番号管理部、２：問合せ要求・結果処理部、３：データベース処理部、４：データベース、１４１：変換テーブル、１４２：文書索引、１４３：索引、１４４：データ、１４５：文書、５：問合せ結果

【図３】

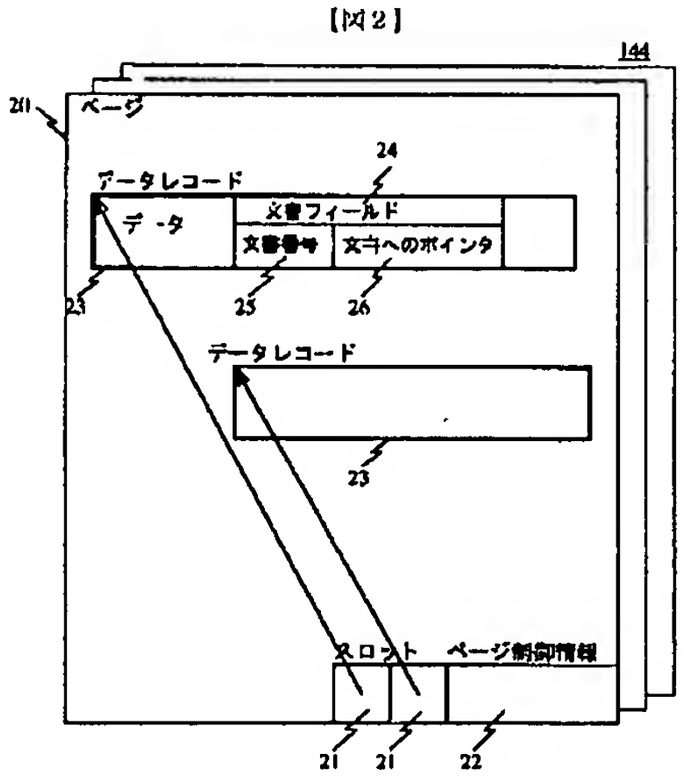
【図3】



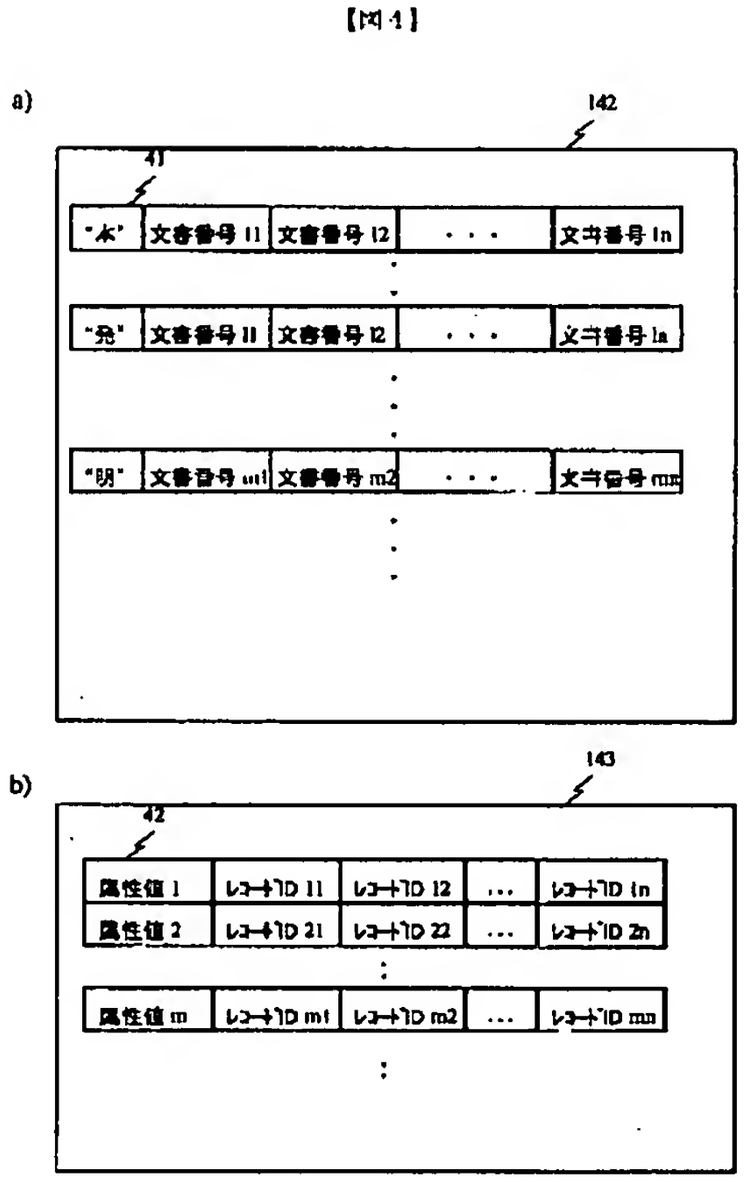
【図１】


【図１】
問合せ要求
1
問合せ結果
5
問合せ要求・結果処理部
問合せ要求受付・解
121
問合せ処理実行制御
122
問合せ結果処理
123
2
データベース処理部
文書索引管理部
索引管理
データ管理部
検索条件合致
文書番号取得
検索条件合致
レコード識別子
取得
134
131
133
文字番号管理部
変換テーブル管理部
文字番号割当
レコード識別子
→文字番号
変換
132
135
3
データベース
4
変換テーブル
文書索引
索引
データ
文書
141
142
143
144
145

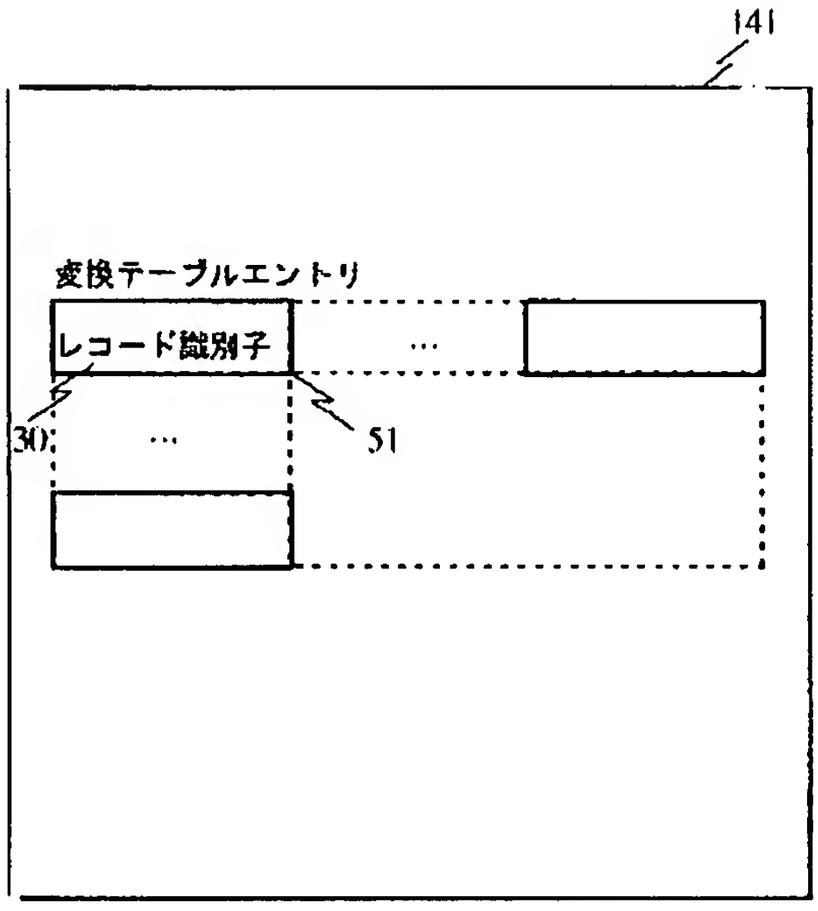
【図２】


【図２】
144
ページ
241
データレコード
24
データ
文書フィールド
文書番号
文字へのポインタ
23
25
26
データレコード
23
スロット
ページ制御情報
21
21
22

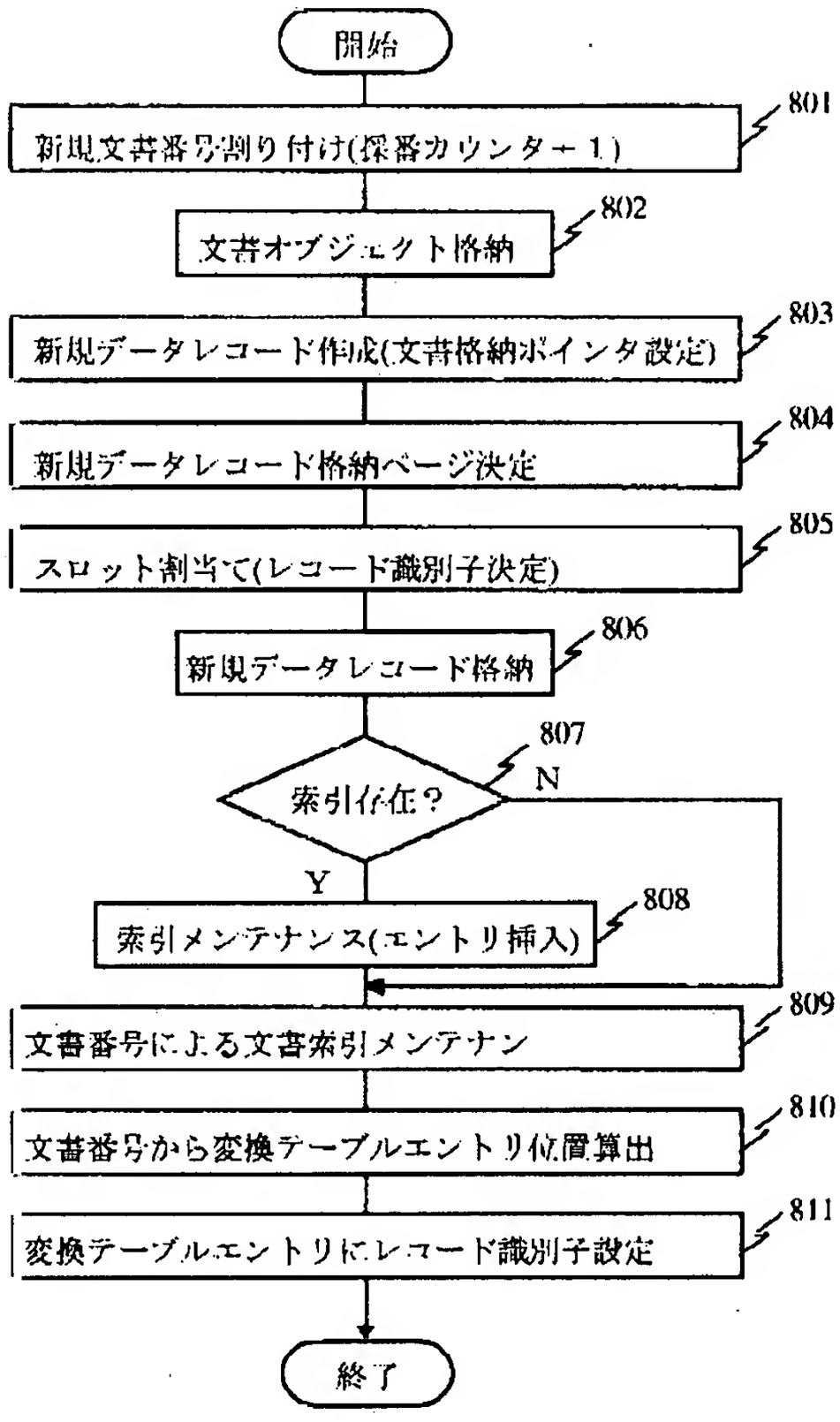
【図４】


【図４】
a)
142
41
"木"
文書番号 t1
文書番号 t2
・・・
文字番号 tn
"発"
文書番号 h1
文書番号 h2
・・・
文字番号 hn
"明"
文書番号 m1
文書番号 m2
・・・
文字番号 mn
b)
143
42
属性値 1
レコードID 11
レコードID 12
・・・
レコードID 1n
属性値 2
レコードID 21
レコードID 22
・・・
レコードID 2n
属性値 m
レコードID m1
レコードID m2
・・・
レコードID mn

【図5】

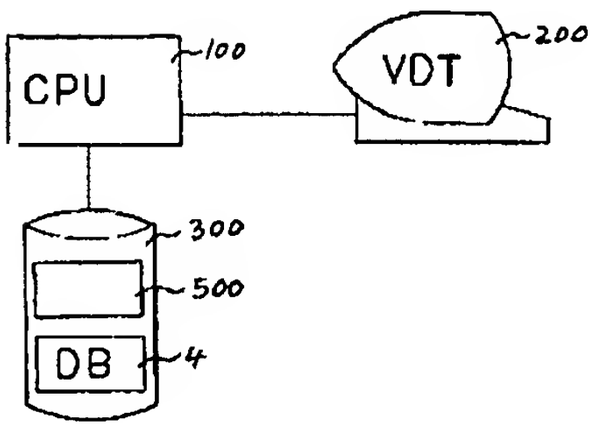

141
変換テーブルエントリ
レコード識別子
…
30
…
51

【図8】


開始
801
新規文書番号割り付け(採番カウンター1)
802
文書オブジェクト格納
803
新規データレコード作成(文書格納ポインタ設定)
804
新規データレコード格納ページ決定
805
スロット割当て(レコード識別子決定)
806
新規データレコード格納
807
索引存在?
N
Y
808
索引メンテナンス(エントリ挿入)
809
文書番号による文書索引メンテナン
810
文書番号から変換テーブルエントリ位置算出
811
変換テーブルエントリにレコード識別子設定
終了

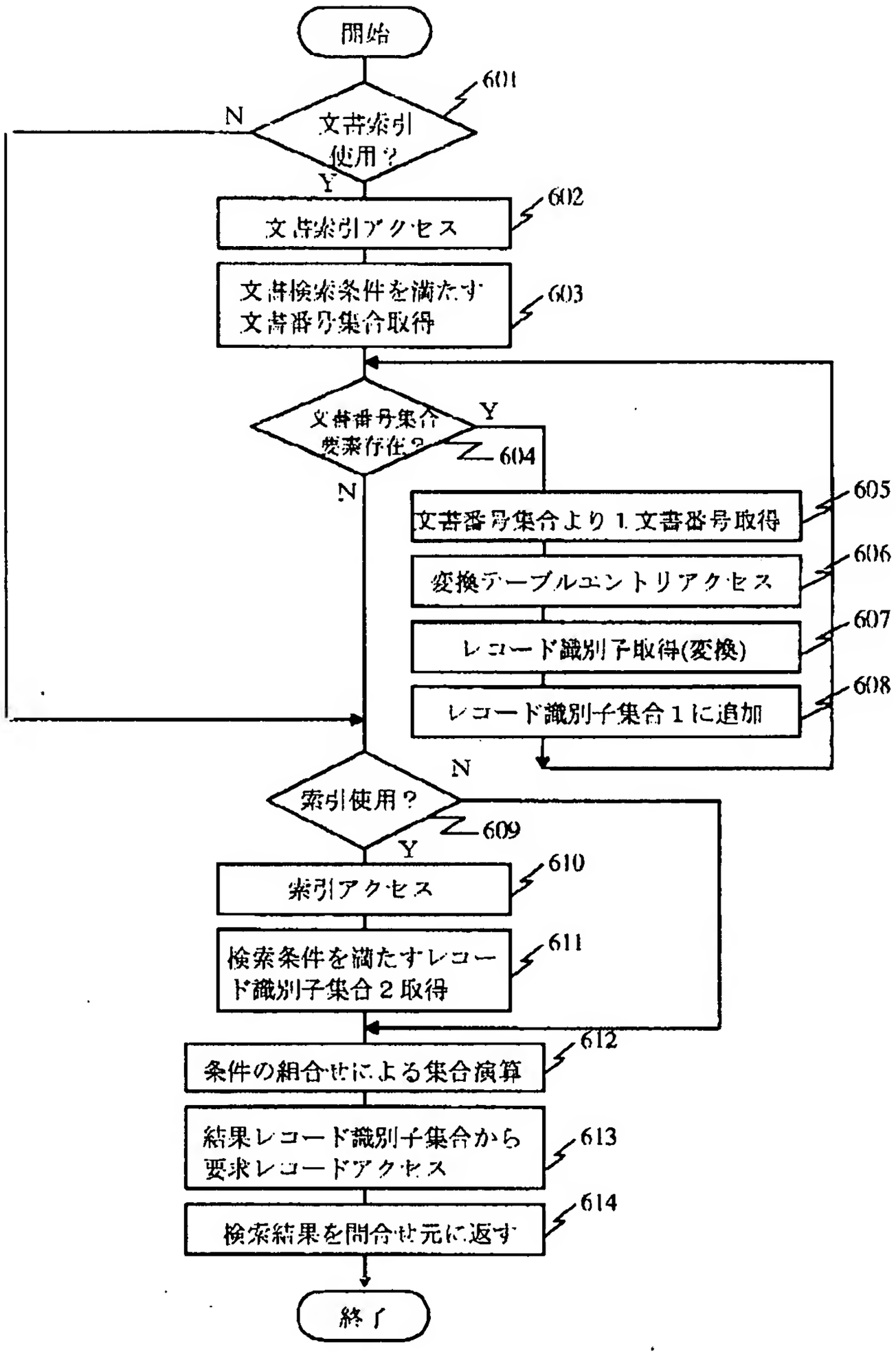
【図14】

図14


CPU
100
VDT
200
300
500
DB
4

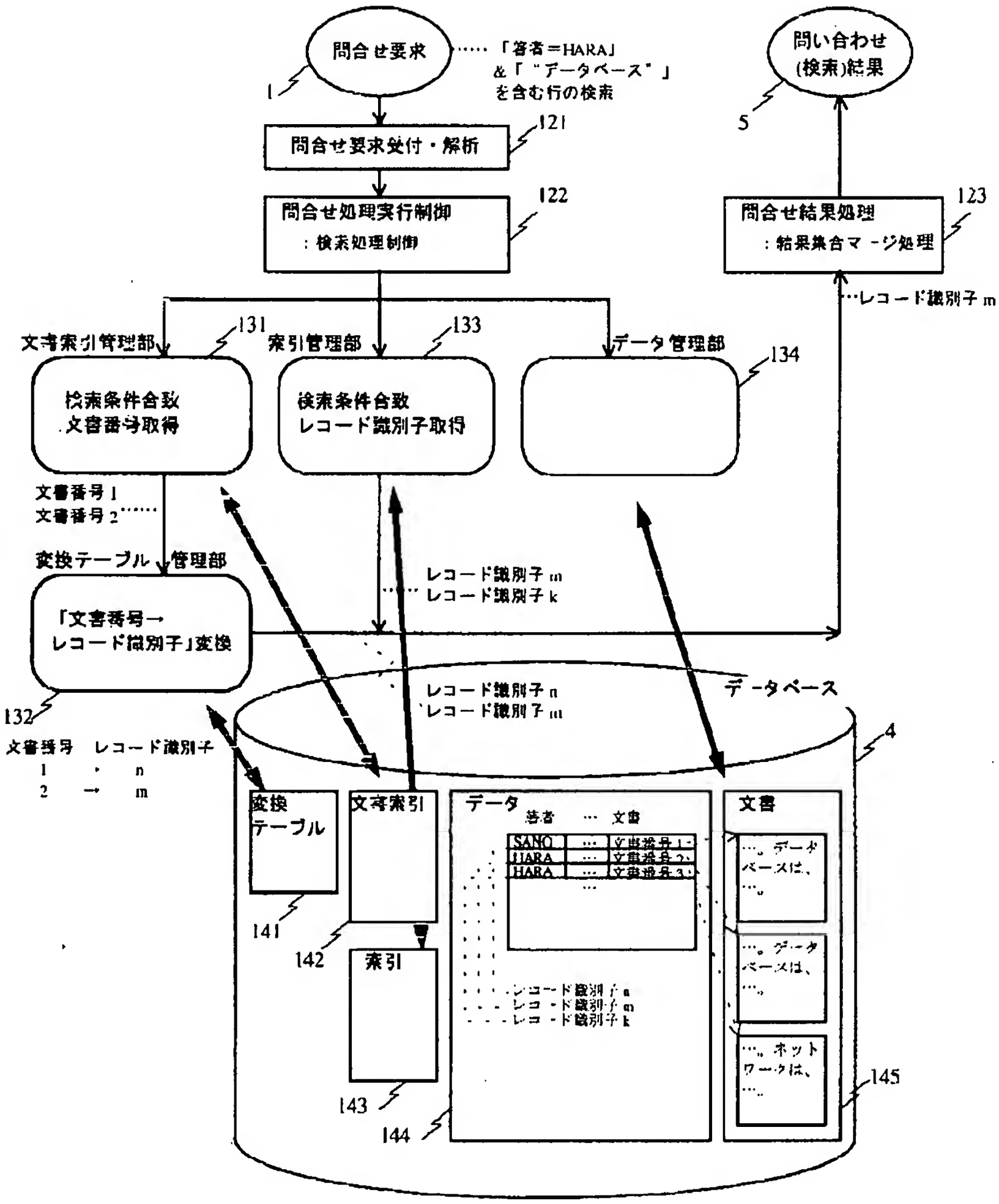
【図6】

【図6】


開始
601
文書索引使用？
N
Y
602
文書索引アクセス
603
文書検索条件を満たす文書番号集合取得
文書番号集合要素存在？
Y
N
604
605
文書番号集合より1文書番号取得
606
変換テーブルエントリアクセス
607
レコード識別子取得(変換)
608
レコード識別子集合1に追加
索引使用？
N
609
Y
610
索引アクセス
611
検索条件を満たすレコード識別子集合2取得
612
条件の組合せによる集合演算
613
結果レコード識別子集合から要求レコードアクセス
614
検索結果を問合せ元に返す
終了

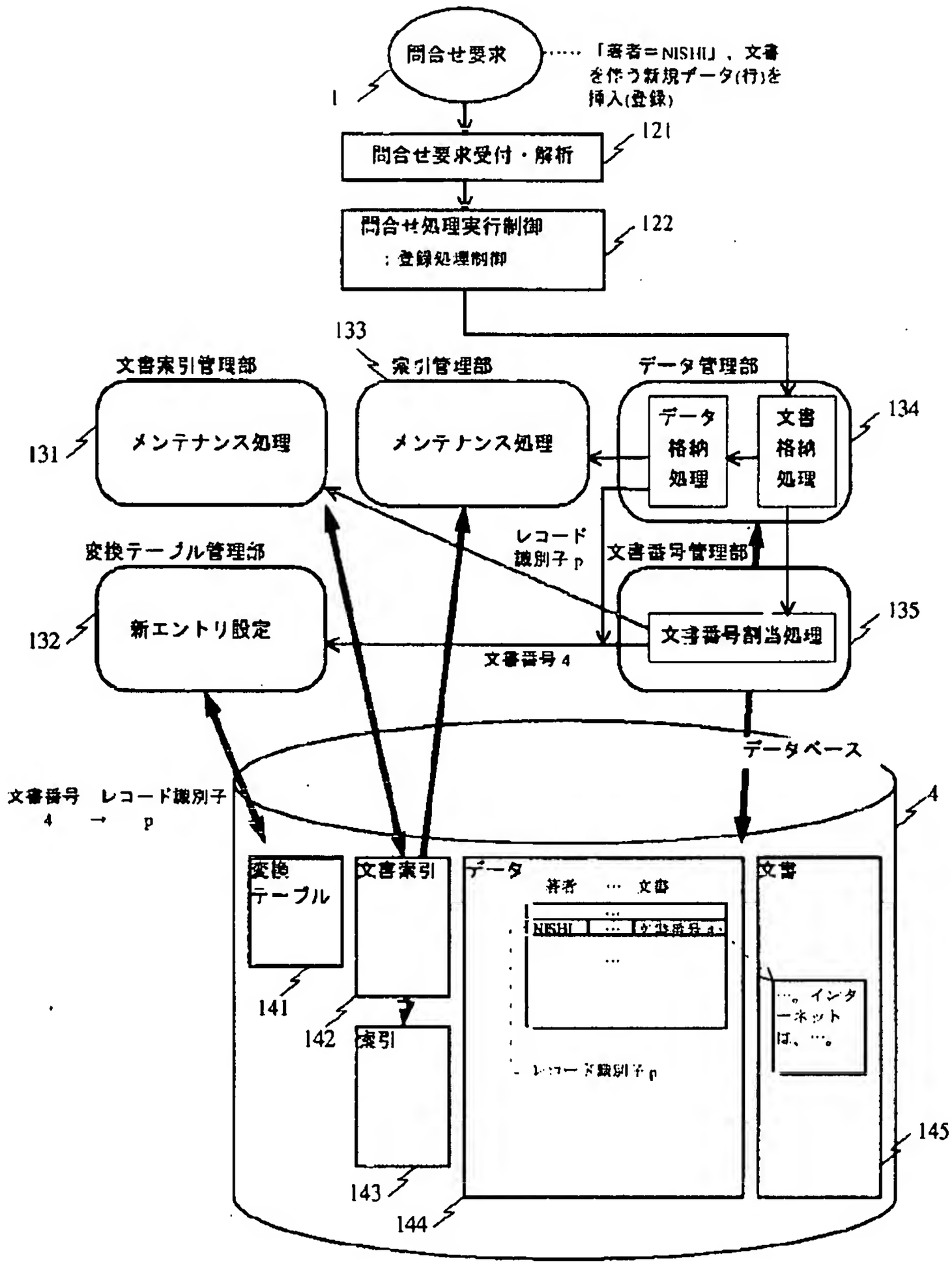
【図７】

【図７】

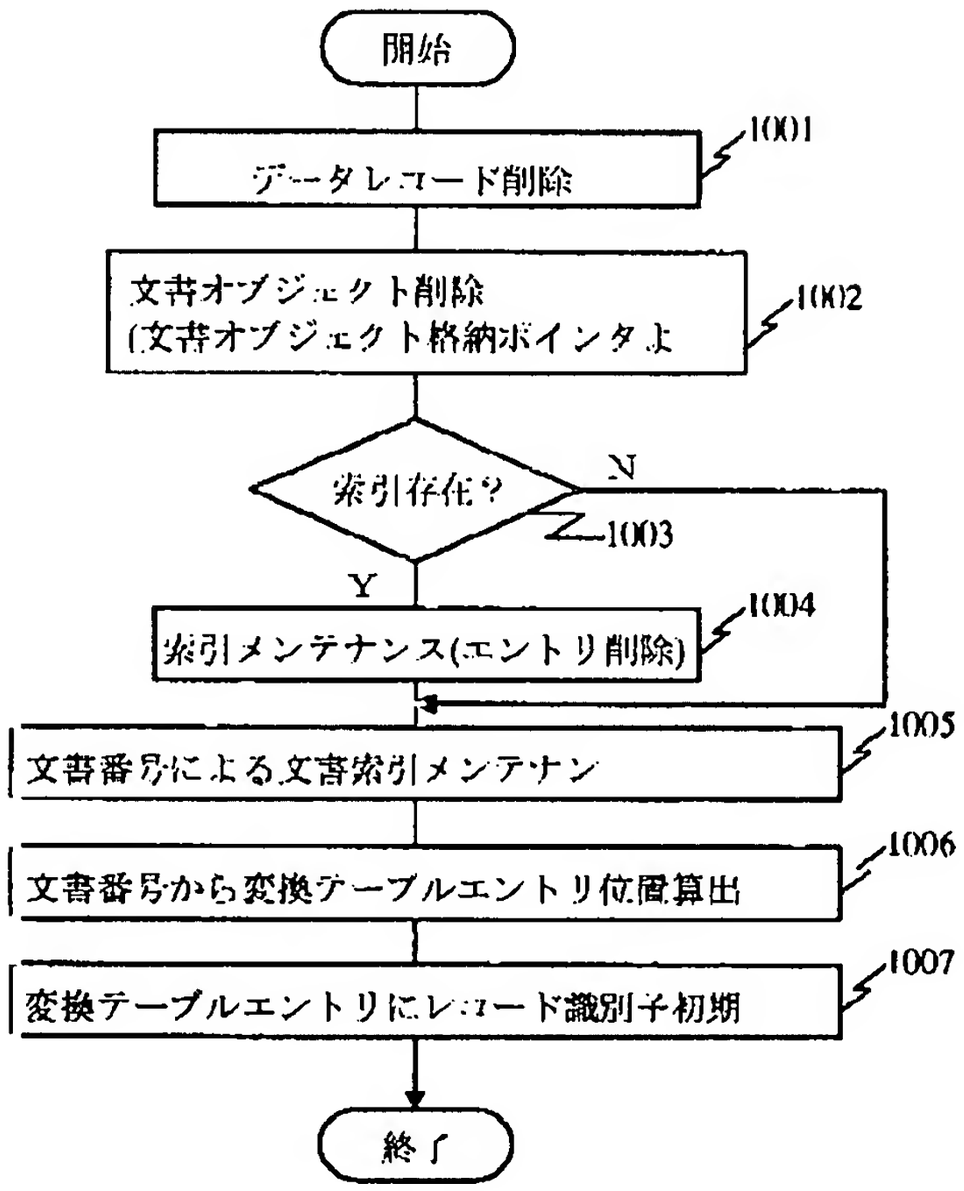

問合せ要求
……「著者＝HARA」
&「"データベース"」
を含む行の検索
1
問合せ要求受付・解析
121
問合せ処理実行制御
：検索処理制御
122
問い合わせ
(検索)結果
5
問合せ結果処理
：結果集合マージ処理
123
…レコード識別子 m
文書索引管理部
131
索引管理部
133
データ管理部
134
検索条件合致
文書番号取得
検索条件合致
レコード識別子取得
文書番号 1
文書番号 2……
変換テーブル管理部
「文書番号→
レコード識別子」変換
132
文書番号 レコード識別子
1 → n
2 → m
レコード識別子 m
…… レコード識別子 k
レコード識別子 n
レコード識別子 m
データベース
4
変換
テーブル
141
文書索引
142
索引
143
データ
著者 … 文書
SANO … 文書番号 1
HARA … 文書番号 2
HARA … 文書番号 3
…
…レコード識別子 n
…レコード識別子 m
…レコード識別子 k
144
文書
…。データベースは、…。
…。データベースは、…。
…。ネットワークは、…。
145

【図9】

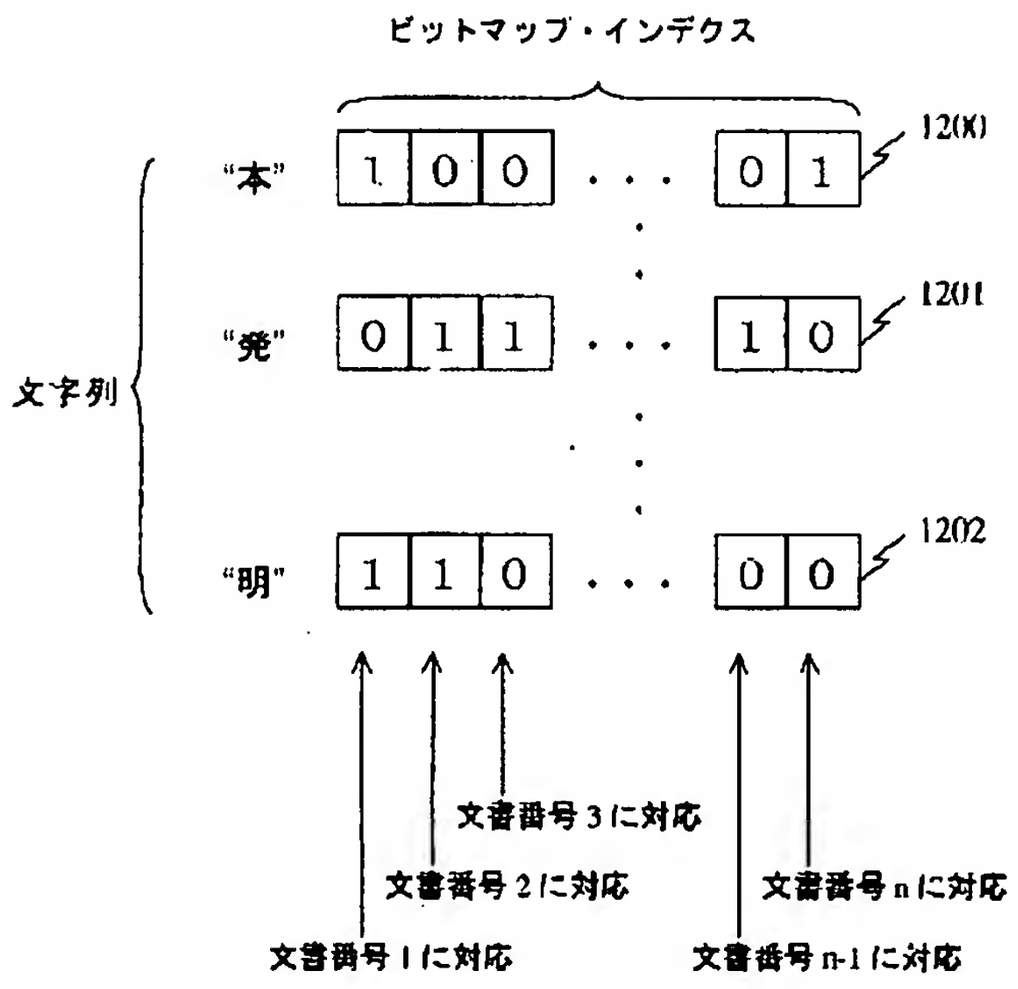
【図9】


問合せ要求
……「著者＝NISHI」，文書を作う新規データ(行)を挿入(登録)
1
問合せ要求受付・解析
121
問合せ処理実行制御
：登録処理制御
122
133
文書索引管理部
索引管理部
データ管理部
131
メンテナンス処理
メンテナンス処理
データ格納処理
文書格納処理
134
変換テーブル管理部
レコード識別子 p
文書番号管理部
132
新エントリ設定
文書番号割当処理
135
文書番号 4
データベース
文書番号　レコード識別子
4　→　p
4
変換テーブル
文書索引
データ
著者　…　文書
NISHI
文書
…。インターネットは、…。
141
142
索引
レコード識別子 p
145
143
144

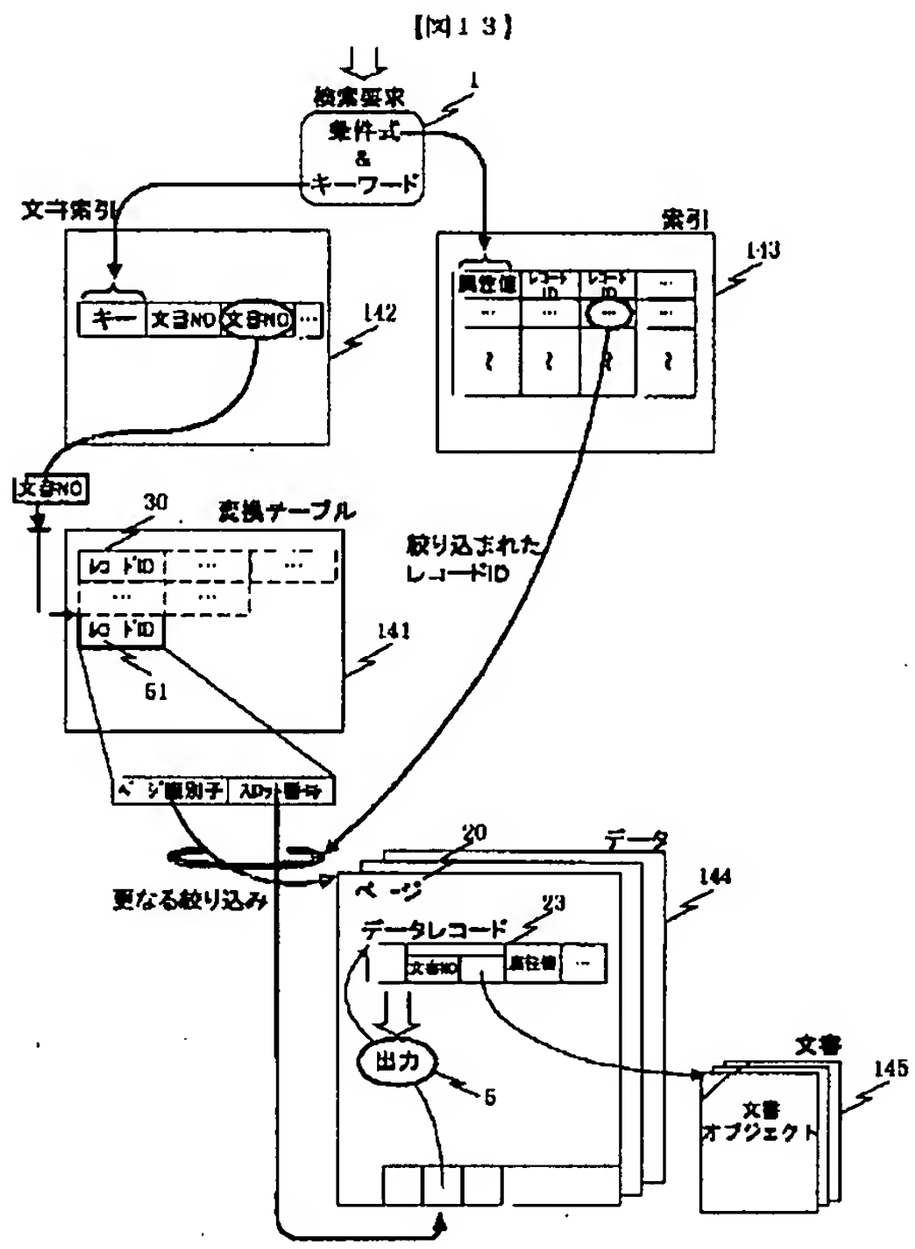
【図１０】


開始
データレコード削除
1001
文書オブジェクト削除
(文書オブジェクト格納ポインタよ
1002
索引存在?
N
1003
Y
索引メンテナンス(エントリ削除)
1004
文書番号による文書索引メンテナン
1005
文書番号から変換テーブルエントリ位置算出
1006
変換テーブルエントリにレコード識別子初期
1007
終了

【図１２】

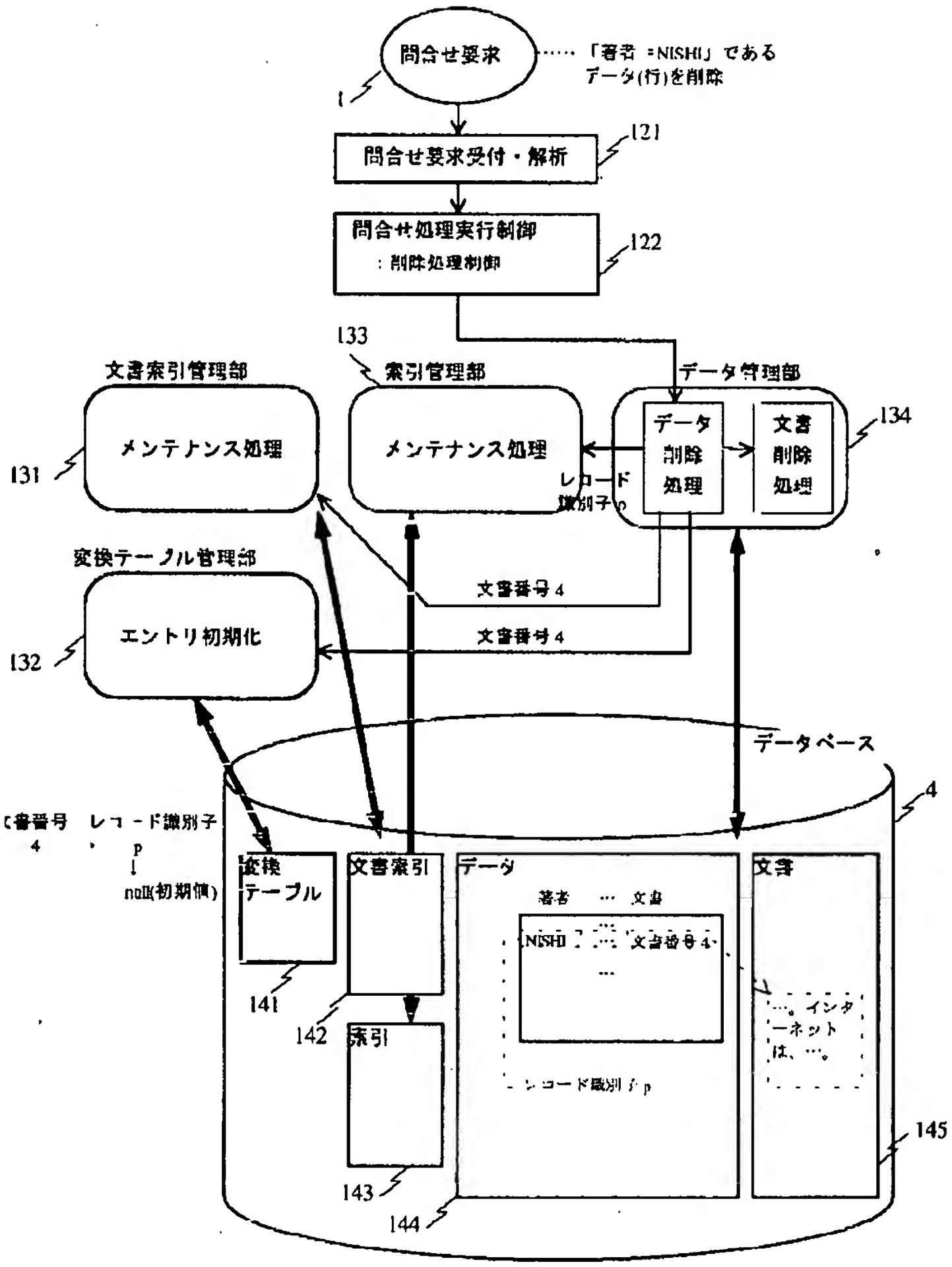

142
ビットマップ・インデクス
"本"
1 0 0 . . . 0 1
1202
"発"
0 1 1 . . . 1 0
1201
文字列
"明"
1 1 0 . . . 0 0
文書番号 3 に対応
文書番号 2 に対応
文書番号 n に対応
文書番号 1 に対応
文書番号 n-1 に対応

【図１３】


検索要求
条件式
&
キーワード
1
文字索引
キー
索引
属性値
レコードID
142
143
文書NO
30
変換テーブル
141
絞り込まれた
レコードID
61
20
データ
ページ
144
データレコード
23
出力
5
文書
オブジェクト
145
更なる絞り込み

【図１１】

【図１１】



フロントページの続き

(72)発明者　河村　信男
神奈川県川崎市幸区鹿島田890番地の12
株式会社日立製作所情報・通信開発本部内

(72)発明者　北村　健一
神奈川県横浜市中区尾上町６丁目81番地
日立ソフトウェアエンジニアリング株式会社内